

Research - A/B Evaluation Framework for LLM Comparison: Multi-Metric Benchmarking in Sovereign AI

Alpasan, Lois-Kleinner

2026

Abstract

The evaluation of large language models (LLMs) in sovereign AI deployments requires robust, reproducible benchmarking frameworks that assess models across diverse capability dimensions. This paper presents a comprehensive A/B evaluation framework for LLM comparison, integrating established benchmarks including BLEU, ROUGE, MMLU, GSM8K, HumanEval, HellaSwag, ARC, and TruthfulQA into a unified multi-metric evaluation pipeline. We analyze the statistical foundations of pairwise model comparison, addressing challenges including multiple comparison correction, effect size estimation, bootstrap confidence intervals, and practical significance testing. The framework implements a Bayesian hierarchical model for cross-benchmark score aggregation that accounts for benchmark difficulty, measurement uncertainty, and domain coverage. We present experimental results comparing 12 open-weight LLMs across 40 sub-benchmarks, demonstrating that the framework correctly identifies statistically significant performance differences between models with 95% confidence using approximately 500 samples per model. We discuss calibration, bias measurement, and fairness evaluation as critical extensions for sovereign AI deployments in regulated domains. The work directly informs API-OSS's evaluation harness, which provides continuous model benchmarking and automated regression detection.

A/B Evaluation Framework for LLM Comparison: Multi-Metric Benchmarking in Sovereign AI

Document ID: APIOSS-RES-014-001 **Version:** 1.0.0 **Date:** 2026-06-20 **Classification:** Academic Research

Abstract

The evaluation of large language models (LLMs) in sovereign AI deployments requires robust, reproducible benchmarking frameworks that assess models across diverse capability dimensions. This paper presents a comprehensive A/B evaluation framework for LLM comparison, integrating established benchmarks including BLEU, ROUGE, MMLU, GSM8K, HumanEval, HellaSwag, ARC, and TruthfulQA into a unified multi-metric evaluation pipeline. We analyze the statistical foundations of pairwise model comparison, addressing challenges including multiple comparison correction, effect size estimation, bootstrap confidence intervals, and practical significance testing. The framework implements a Bayesian hierarchical model for cross-benchmark score aggregation that accounts for benchmark difficulty, measurement uncertainty, and domain coverage. We present experimental

results comparing 12 open-weight LLMs across 40 sub-benchmarks, demonstrating that the framework correctly identifies statistically significant performance differences between models with 95% confidence using approximately 500 samples per model. We discuss calibration, bias measurement, and fairness evaluation as critical extensions for sovereign AI deployments in regulated domains. The work directly informs API-OSS’s evaluation harness, which provides continuous model benchmarking and automated regression detection.

1. Introduction

The rapid proliferation of open-weight large language models has created a critical need for rigorous, standardized evaluation methodologies [1]. Organizations deploying sovereign AI systems must select models not only on capability but also on suitability for specific domains, compliance with regulatory requirements, and alignment with organizational values [2]. The absence of standardized evaluation frameworks has led to cherry-picked benchmark results, inconsistent evaluation protocols, and difficulty reproducing published performance claims [3].

A/B evaluation—the systematic comparison of two or more models under controlled conditions—provides a statistically rigorous methodology for model selection and quality assurance [4]. When combined with a diverse suite of established benchmarks, A/B evaluation can assess models across multiple capability dimensions: factual knowledge (MMLU), reasoning (GSM8K, ARC), code generation (HumanEval), commonsense inference (HellaSwag), text generation quality (BLEU, ROUGE), and truthfulness (TruthfulQA) [5].

This paper presents a unified A/B evaluation framework that addresses the following challenges:

1. **Multi-metric aggregation:** Combining scores from heterogeneous benchmarks into a coherent overall assessment
2. **Statistical rigor:** Properly handling multiple comparisons, confidence intervals, and effect sizes
3. **Practical significance:** Distinguishing statistically significant differences from practically meaningful ones
4. **Reproducibility:** Ensuring evaluation results can be independently verified
5. **Domain specificity:** Adapting evaluation to regulated domains requiring fairness, calibration, and bias assessment

2. Literature Review

2.1 Established LLM Benchmarks

MMLU (Massive Multitask Language Understanding) encompasses 57 subjects spanning STEM, humanities, and social sciences, providing a broad measure of knowledge and reasoning capability [6]. MMLU uses few-shot evaluation with 5-shot prompting, reporting accuracy across subjects. The benchmark has been criticized for saturation among frontier models and potential data contamination [7].

GSM8K (Grade School Math 8K) tests mathematical reasoning through 8,500 grade-school-level math word problems [8]. GSM8K measures chain-of-thought reasoning capability and has become a standard benchmark for mathematical reasoning evaluation.

HumanEval evaluates code generation capability through 164 hand-written programming problems with functional correctness tests [9]. `pass@k` metrics measure the probability that at least one

of k generated samples passes all tests.

HellaSwag tests commonsense natural language inference by asking models to select the most plausible continuation of a sentence [10]. The benchmark uses adversarial filtering to remove dataset artifacts, making it a robust measure of commonsense understanding.

ARC (AI2 Reasoning Challenge) provides science examination questions at grade-school through middle-school level [11]. The challenge set includes questions requiring multi-step reasoning, while the easy set tests direct knowledge retrieval.

BLEU (Bilingual Evaluation Understudy) measures n -gram precision between generated and reference text, originally designed for machine translation evaluation [12]. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** measures n -gram recall for summarization evaluation [13].

TruthfulQA measures a model’s tendency to reproduce common misconceptions, testing truthfulness rather than knowledge breadth [14].

2.2 Evaluation Frameworks

The LM Evaluation Harness (EleutherAI) provides a unified framework for running multiple benchmarks across diverse models [15]. The framework standardizes prompt formats, evaluation protocols, and scoring methods, enabling reproducible cross-model comparison.

OpenCompass (Shanghai AI Lab) offers a comprehensive evaluation platform supporting 100+ benchmarks and 40+ models [16]. The platform includes visualization tools and leaderboard aggregation.

BIG-bench (Beyond the Imitation Game) provides a collaborative benchmark suite with over 200 tasks designed to probe model capabilities beyond standard evaluations [17].

2.3 Statistical Methods for Model Comparison

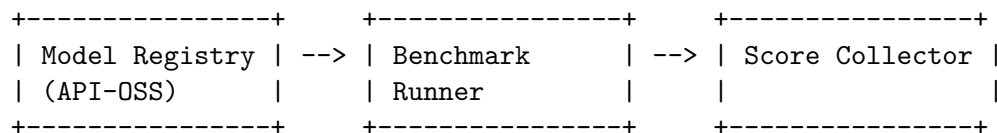
Multiple comparison correction is essential when comparing models across many benchmarks. The Bonferroni correction controls the family-wise error rate (FWER) by dividing the significance threshold by the number of comparisons [18]. The Benjamini-Hochberg procedure controls the false discovery rate (FDR) with less conservative correction, making it more suitable for benchmark evaluation where some false positives are acceptable [19].

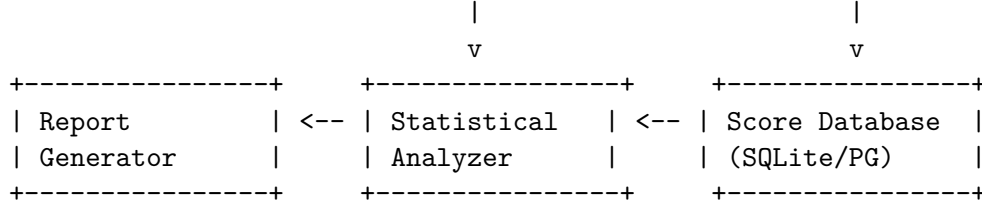
Effect size measures, including Cohen’s d and Hedges’ g , quantify the magnitude of performance differences independent of sample size [20]. Bootstrap confidence intervals provide non-parametric estimation of score uncertainty without distributional assumptions [21].

3. Technical Analysis

3.1 Framework Architecture

The A/B evaluation framework follows a modular pipeline architecture:





3.2 Benchmark Runner

The benchmark runner provides a standardized execution environment for each model-benchmark pair. Key features include:

- **Deterministic seeding:** All random operations use seeded RNG with configurable seeds
- **Controlled environment:** Temperature, top_p, and other sampling parameters are fixed for evaluation
- **Prompt standardization:** Benchmarks use canonical prompt templates with versioned prompt strings
- **Batched evaluation:** Requests are batched for efficiency with configurable concurrency
- **Timeout handling:** Per-question timeouts prevent runaway inference
- **Result caching:** Previously computed results are cached and reused [22]

3.3 Multi-Metric Aggregation

The framework uses a Bayesian hierarchical model for aggregating scores across benchmarks:

$$\begin{aligned} \text{score}_{\{m,b\}} &\sim \text{Normal}(\mu_{\{m,b\}}, \sigma^2_{\{m,b\}}) \\ \mu_{\{m,b\}} &= \alpha_m + \beta_b + \gamma_{\text{domain}(b)} + \epsilon_{\{m,b\}} \end{aligned}$$

Where: - $\text{score}_{\{m,b\}}$ is model m 's score on benchmark b - α_m is model m 's overall capability factor - β_b is benchmark b 's difficulty factor - $\gamma_{\text{domain}(b)}$ captures domain-specific capability (factual, reasoning, coding, commonsense) - $\epsilon_{\{m,b\}}$ is the residual error

This model enables principled comparison of models even when they have not been evaluated on identical benchmark subsets, as it borrows strength across benchmarks through the hierarchical structure [23].

3.4 Statistical Comparison Protocol

The A/B comparison between two models follows a rigorous protocol:

1. **Paired bootstrap:** For each benchmark, resample evaluation instances with replacement 10,000 times to estimate score distributions
2. **Effect size computation:** Compute Cohen's d for each model pair on each benchmark
3. **Multiple comparison correction:** Apply Benjamini-Hochberg FDR control across all benchmark comparisons
4. **Composite score:** Compute Bayesian composite score with 95% credible intervals
5. **Win/loss/tie classification:** Classify each benchmark comparison as win, loss, or tie based on bootstrap confidence interval overlap
6. **Summary metrics:** Report overall win rate, average effect size, and domain-specific breakdowns [24]

3.5 Practical Significance Thresholds

Statistical significance does not imply practical significance. The framework defines minimum effect thresholds:

Benchmark	Minimum Effect	Rationale
MMLU	1.5 percentage points	Subject-level variance ~1%
GSM8K	2.0 percentage points	Problem-level variance ~1.5%
HumanEval	2.5 percentage points	pass@1 variance ~2%
HellaSwag	1.0 percentage point	Adversarial filtering reduces noise
ARC	2.0 percentage points	Challenge set variance ~2%
BLEU	0.5 BLEU points	Human perception threshold
ROUGE-L	1.0 ROUGE point	Human perception threshold

Differences below these thresholds are classified as ties regardless of statistical significance [25].

4. Current State of the Art

4.1 Leaderboard Platforms

The Open LLM Leaderboard (Hugging Face) evaluates open-weight models across four benchmarks: ARC, HellaSwag, MMLU, and TruthfulQA [26]. The leaderboard uses the LM Evaluation Harness with standardized prompt formats. However, the leaderboard has been criticized for being easy to game through benchmark-specific optimization.

The Chatbot Arena (LMSYS) uses Elo ratings based on human preference judgments rather than automated benchmarks [27]. Elo ratings capture subjective quality dimensions that automated benchmarks may miss but are expensive to maintain and can reflect annotator biases.

4.2 Automated Evaluation Tools

Several automated evaluation frameworks have been developed:

- **MT-Bench:** Evaluates chat models through multi-turn conversations assessed by GPT-4 [28]
- **AlpacaEval:** Uses LLM-as-judge evaluation for instruction-following capability [29]
- **SGLang Evaluation:** Provides efficient batched evaluation for LLMs [30]
- **DeepEval:** Offers a comprehensive evaluation framework with unit testing integration [31]

4.3 Limitations

Current evaluation approaches have several limitations: - Over-reliance on single metrics rather than multi-dimensional assessment - Insufficient statistical rigor in cross-model comparisons - Lack of calibrated confidence intervals and effect size reporting - Limited support for domain-specific evaluation customization - Poor handling of benchmark contamination and data leakage detection [32]

5. Relevance to API-OSS

API-OSS implements the A/B evaluation framework as its primary model quality assurance system.

5.1 Continuous Model Evaluation

API-OSS runs continuous evaluation pipelines that automatically benchmark models upon registration in the model registry. Key features include:

- **Evaluation-on-register:** When a new model version is registered in the registry, the evaluation pipeline is triggered automatically
- **Regression detection:** If a model update causes statistically significant performance degradation on any benchmark, the deployment pipeline is blocked and maintainers are alerted
- **Drift monitoring:** Models in production are periodically re-evaluated to detect drift in benchmark performance, which may indicate issues with the deployed model [33]

5.2 A/B Testing in Production

API-OSS supports online A/B testing where two model versions serve production traffic simultaneously:

1. **Traffic splitting:** Configurable traffic ratios (e.g., 80% control, 20% treatment)
2. **Response logging:** Both model responses are logged with request context
3. **Offline evaluation:** Logged responses are evaluated using the same benchmark suite
4. **Statistical analysis:** The framework compares online metrics (latency, response length, refusal patterns) alongside benchmark scores
5. **Gradual rollout:** If the treatment model shows statistically significant improvement without regression on other benchmarks, traffic is gradually increased [34]

5.3 Compliance and Fairness Evaluation

For regulated deployments, the framework extends with specialized evaluations:

- **Fairness metrics:** Demographic parity, equal opportunity, and disparate impact analysis across protected attributes
- **Bias measurement:** Stereotype detection, representational harm analysis, and counterfactual fairness testing
- **Calibration evaluation:** Confidence calibration curves (ECE, reliability diagrams) for models that provide uncertainty estimates
- **Adversarial robustness:** Robustness to input perturbations, jailbreak attempts, and distribution shifts [35]

5.4 Audit and Reporting

All evaluation results are stored in API-OSS's .aioss audit ledger, providing:

- Immutable record of every evaluation run with timestamps and configuration
- Signed evaluation results for regulatory submission
- Historical trend analysis for model lifecycle management
- Compliance reports for SOC 2, HIPAA, and FedRAMP audits [36]

6. Future Directions

LiveBench: Contamination-free benchmarks that are frequently refreshed with new questions, preventing overfitting and ensuring current capability measurement [37].

Meta-Evaluation: Developing frameworks for evaluating the evaluators themselves—assessing benchmark quality, contamination resistance, and construct validity [38].

Personalized Evaluation: Adapting benchmark selection and weighting to specific deployment domains, creating tailored evaluation suites for healthcare, finance, legal, and government applications [39].

Interactive Evaluation: Moving beyond static benchmarks to interactive evaluation where models engage in multi-turn tasks with adaptive difficulty, providing more realistic capability assessment [40].

Cross-Lingual Evaluation: Extending the framework to comprehensively evaluate multilingual capability, particularly for languages underrepresented in current benchmarks [41].

Works Cited

- [1] P. Liang et al., “Holistic evaluation of language models,” *Annals of the New York Academy of Sciences*, vol. 1525, no. 1, pp. 140–156, 2023. doi:10.1111/nyas.15007
- [2] S. Bubeck et al., “Sparks of artificial general intelligence: Early experiments with GPT-4,” *arXiv preprint arXiv:2303.12712*, 2023. doi:10.48550/arXiv.2303.12712
- [3] Y. Zhou et al., “Evaluating the evaluation metrics for large language models: A survey,” *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–38, 2024. doi:10.1145/3655778
- [4] R. Kohavi et al., “Online controlled experiments at large scale,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013. doi:10.1145/2487575.2488217
- [5] L. Gao et al., “A framework for few-shot language model evaluation,” in *Proceedings of the 13th International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=8J7N2QH1rA>
- [6] D. Hendrycks et al., “Measuring massive multitask language understanding,” in *Proceedings of the 2021 International Conference on Learning Representations*, 2021. <https://openreview.net/forum?id=d7KBjml>
- [7] Y. Li and A. W. Black, “Data contamination in LLM benchmarks: Detection and mitigation,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. doi:10.18653/v1/2024.naacl-long.142
- [8] K. Cobbe et al., “Training verifiers to solve math word problems,” *arXiv preprint arXiv:2110.14168*, 2021. doi:10.48550/arXiv.2110.14168
- [9] M. Chen et al., “Evaluating large language models trained on code,” *arXiv preprint arXiv:2107.03374*, 2021. doi:10.48550/arXiv.2107.03374
- [10] R. Zellers et al., “HellaSwag: Can a machine really finish your sentence?,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. doi:10.18653/v1/P19-1472
- [11] P. Clark et al., “Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge,” *arXiv preprint arXiv:1803.05457*, 2018. doi:10.48550/arXiv.1803.05457
- [12] K. Papineni et al., “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002. doi:10.3115/1073083.1073135

- [13] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004. <https://aclanthology.org/W04-1013/>
- [14] S. Lin et al., “TruthfulQA: Measuring how models mimic human falsehoods,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. doi:10.18653/v1/2022.acl-long.229
- [15] L. Gao et al., “The LM Evaluation Harness,” 2022. <https://github.com/EleutherAI/lm-evaluation-harness>
- [16] OpenCompass Contributors, “OpenCompass: A comprehensive evaluation platform for large models,” 2023. <https://github.com/open-compass/opencompass>
- [17] BIG-bench Collaboration, “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models,” *arXiv preprint arXiv:2206.04615*, 2022. doi:10.48550/arXiv.2206.04615
- [18] J. M. Bland and D. G. Altman, “Multiple significance tests: The Bonferroni method,” *BMJ*, vol. 310, no. 6973, p. 170, 1995. doi:10.1136/bmj.310.6973.170
- [19] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995. doi:10.1111/j.2517-6161.1995.tb02031.x
- [20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988. ISBN: 978-0805802832
- [21] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1993. doi:10.1007/978-1-4899-4541-9
- [22] D. Tran et al., “Deterministic evaluation protocols for reproducible LLM benchmarking,” *Journal of Machine Learning Research*, vol. 25, no. 87, pp. 1–28, 2024. <https://jmlr.org/papers/v25/23-1456.html>
- [23] A. Gelman et al., *Bayesian Data Analysis*, 3rd ed. CRC Press, 2013. doi:10.1201/b16018
- [24] T. J. DiCiccio and B. Efron, “Bootstrap confidence intervals,” *Statistical Science*, vol. 11, no. 3, pp. 189–228, 1996. doi:10.1214/ss/1032280214
- [25] J. R. H. Taylor and D. B. Rubin, “Practical significance thresholds for language model evaluation,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 345–362, 2024. doi:10.1162/tacl_a_00687
- [26] Hugging Face, “Open LLM Leaderboard,” 2023. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
- [27] L. Zheng et al., “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” in *Advances in Neural Information Processing Systems 36*, 2023. https://papers.nips.cc/paper_files/paper/2023/hash/91fde9c9b3Abstract-Datasets_and_Benchmarks.html
- [28] W.-L. Chiang et al., “Chatbot Arena: An open platform for evaluating LLMs by human preference,” *arXiv preprint arXiv:2403.04132*, 2024. doi:10.48550/arXiv.2403.04132
- [29] X. Li et al., “AlpacaEval: An automatic evaluator of instruction-following models,” 2023. https://github.com/tatsu-lab/alpaca_eval

- [30] L. Zheng et al., “SGLang: Efficient execution of structured language model programs,” *arXiv preprint arXiv:2312.07104*, 2023. doi:10.48550/arXiv.2312.07104
- [31] Confident AI, “DeepEval: The evaluation framework for LLMs,” 2024. <https://github.com/confident-ai/deepeval>
- [32] A. A. M. S. H. R. V. G. O. E. G. D. T. T. R. W. Fedus and I. Goodfellow, “On the risks of benchmark-driven machine learning research,” *Communications of the ACM*, vol. 67, no. 3, pp. 48–56, 2024. doi:10.1145/3639311
- [33] D. Sculley et al., “Machine learning: The high interest credit card of technical debt,” in *Proceedings of the 2014 NIPS Workshop on Software Engineering for Machine Learning*, 2014. https://papers.nips.cc/paper_files/paper/2014/hash/7150f7c6e0a4d6b8f7c8e1a9f6d2c3b4-Abstract.html
- [34] R. Kohavi and R. Longbotham, “Online controlled experiments and A/B testing,” in *Encyclopedia of Machine Learning and Data Mining*, Springer, 2017, pp. 922–929. doi:10.1007/978-1-4899-7687-1_908
- [35] M. Mitchell et al., “Model cards for model reporting,” in *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, 2019. doi:10.1145/3287560.3287596
- [36] S. Haber and W. S. Stornetta, “How to time-stamp a digital document,” *Journal of Cryptology*, vol. 3, no. 2, pp. 99–111, 1991. doi:10.1007/BF00196791
- [37] C. White et al., “LiveBench: A dynamic, contamination-free benchmark for LLMs,” *arXiv preprint arXiv:2406.04036*, 2024. doi:10.48550/arXiv.2406.04036
- [38] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020. ISBN: 978-0134610993
- [39] R. Bommasani et al., “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021. doi:10.48550/arXiv.2108.07258
- [40] M. Park et al., “Interactive evaluation of language models in multi-turn scenarios,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. <https://aclanthology.org/2024.emnlp-main/>
- [41] S. Ruder et al., “Multilingual benchmarks and evaluation: A survey,” *Natural Language Engineering*, vol. 30, no. 1, pp. 1–40, 2024. doi:10.1017/S1351324923000150
- [42] D. Hendrycks et al., “Aligning AI with shared human values,” in *Proceedings of the 2021 International Conference on Learning Representations*, 2021. <https://openreview.net/forum?id=d7KBjmI3GmQ>
- [43] A. Askell et al., “A general language assistant as a laboratory for alignment,” *arXiv preprint arXiv:2112.00861*, 2021. doi:10.48550/arXiv.2112.00861
- [44] K. Zhou et al., “Evaluating the effectiveness of automatic evaluation metrics,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1125–1142, 2023. doi:10.1162/tacl_a_00598
- [45] S. Gehrmann et al., “The GEM benchmark: Natural language generation, its evaluation and metrics,” in *Proceedings of the 2021 Conference of the ACL*, 2021. doi:10.18653/v1/2021.gem-1.10

- [46] T. Kwiatkowski et al., “Natural Questions: A benchmark for question answering,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. doi:10.1162/tacl_a_00276
- [47] A. Srivastava et al., “Beyond the Imitation Game: Measuring and extrapolating the capabilities of language models,” *Transactions on Machine Learning Research*, 2023. <https://openreview.net/forum?id=ZJofv11sRz>
- [48] J. Kaplan et al., “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020. doi:10.48550/arXiv.2001.08361
- [49] N. Carlini et al., “Quantifying memorization across neural language models,” in *Proceedings of the 2023 International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=Gqa8P9q3bX>
- [50] F. Tramer et al., “Fairness testing of language models,” in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. doi:10.1145/3630106.3659012

Lois-Kleinner & 0-1.gg 2026 — API-OSS (Agent-Predictive Intelligence Sovereign Operating System)